Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides

Associations Between Genome-wide Gene Expression and Ambient Nitrogen Oxides

Nahid Mostafavi,[a] *Jelle Vlaanderen,*[a] *Lutzen Portengen,*[a] *Marc Chadeau-Hyam,*[a,b] *Lars Modig,*[c] *Domenico Palli,*[e] *Ingvar A. Bergdahl,*[c,d] *Bert Brunekreef,*[a,f] *Paolo Vineis,*[b,g] *Dennie G. A. J. Hebels,*[h] *Jos C. S. Kleinjans,*[h] *Vittorio Krogh,*[i] *Gerard Hoek,*[a] *Panagiotis Georgiadis,*[j] *Soterios A. Kyrtopoulos,*[j] *and Roel Vermeulen*[a,b]

Background: We hypothesize that biological perturbations due to exposure to ambient air pollution are reflected in gene expression levels in peripheral blood mononuclear cells.

Methods: We assessed the association between exposure to ambient air pollution and genome-wide gene expression levels in peripheral blood mononuclear cells collected from 550 healthy subjects participating in cohorts from Italy and Sweden. Annual air pollution estimates of nitrogen oxides (NO_x) at time of blood collection (1990–2006) were available from the ESCAPE study. In addition to univariate analysis and two variable selection methods to investigate the association between expression and exposure to NO_x, we applied gene set enrichment analysis to assess overlap between our most perturbed genes and gene sets hypothesized to be related to air pollution and cigarette smoking. Finally, we assessed associations between NO_x and CpG island methylation at the identified genes.

Results: Annual average NO_x exposure in the Italian and Swedish cohorts was 94.2 and 6.7 μg/m^3, respectively. Long-term exposure to NO_x was associated with seven probes in the Italian cohort and one probe in the Swedish (and combined) cohorts. For genes *AHCYL2* and *MTMR2*, changes were also seen in the methylome. Genes hypothesized to be downregulated due to cigarette smoking were enriched among the most strongly downregulated genes from our study.

Conclusion: This study provides evidence of subtle changes in gene expression related to exposure to long-term NO_x. On a global level, the observed changes in the transcriptome may indicate similarities between air pollution and tobacco induced changes in the transcriptome.

(*Epidemiology* 2017;28: 320–328)

Epidemiologic studies have consistently shown associations between long-term exposure to ambient air pollution and incidence and prevalence of chronic diseases, such as respiratory and cardiovascular disease.[1,2] Although the exact mechanisms responsible for these adverse health effects are unclear, several studies have suggested pollutant-induced oxidative stress and systemic inflammation as potential intermediate biological responses to air pollutants.[3–5] It has been hypothesized that (early) systemic effects of long-term exposure to air pollution can be detected by assessing genome-wide gene expression profiles in peripheral blood mononuclear cells.[6,7]

Although exposure to air pollutants has been shown to induce changes in gene expression in animal and in vitro experiments,[8–10] evidence from human studies is scarce.[11,12] Relevant evidence comes from a study by Wittkopp et al.[13] In this panel study, week-long exposure assessment of ambient air pollutants was combined with the assessment of expression levels of 35 candidate genes from 10 biologic pathways relevant to air pollution exposure responses.[13] Positive associations were observed between traffic-related pollutants elemental carbon, black carbon, primary organic carbon, polycyclic aromatic hydrocarbons (PAHs) in particulate matter (PM), and nitrogen oxides (NO_x) and the *Nrf2* gene (*NFE2L2*), and *Nrf2*-mediated genes, *HMOX1*, *NQO1*, and *SOD2*.

Submitted 8 December 2015; accepted 19 January 2017.

From the [a]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands; [b]Medical Research Council-Health Protection Agency Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom; [c]Department of Public Health and Clinical Medicine, Occupational and Environmental Medicine, Umeå University, Umeå, Sweden; [d]Department of Biobank Research, Umeå University, Umeå, Sweden; [e]Molecular and Nutritional Epidemiology Unit, Cancer Prevention and Research Institute (ISPO), Florence, Italy; [f]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; [g]HuGeF Foundation, Turin, Italy; [h]Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands; [i]Epidemiology Unit, Instituto Tumori, Milan, Italy; and [j]National Hellenic Research Foundation, Institute of Biology, Pharmaceutical Chemistry and Biotechnology, Athens, Greece.

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007e2011) under Grant Agreement Number: 211250 (the European Study of Cohorts for Air Pollution Effects), 226756 (Envirogenomarkers), and 308610 (Exposomics). Data of EnviroGenoMarkers will not be made available online, because the use of these samples for anything other than we have ethical permission for is prohibited. The results of this study are available for re-evaluation whenever needed.

The authors report no conflicts of interest.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Roel Vermeulen, Division Environmental Epidemiology, Institute for Risk Assessment Sciences, Yalelaan 2, Room 353, 3584 CM, Utrecht, The Netherlands. E-mail: r.c.h.vermeulen@uu.nl.

ISSN: 1044-3983/17/2803-0320
DOI: 10.1097/EDE.0000000000000628

Three short-term (1–2 hour) inhalation studies reported changes in the expression of genes involved in inflammation, tissue growth, and host defense, including *IGF-1* signaling, insulin receptor signaling, and *NRF2*-mediated oxidative stress response pathway in response to exposure to ultrafine particles[6]; increased expression levels of genes involved in vascular inflammation and hemostasis (e.g., *IL8RA*, *TNFAIP6*, and *VEGF*) in response to 2 hours exposure to diesel exhaust[14]; and genes involved in oxidative stress, protein degradation, and coagulation (e.g., *PLAU*, *F2R*, *CBL*, *UBR1*) in response to 1 hour of exposure to diesel exhaust.[7] To date, studies are largely inconclusive and have not resulted in clear gene expression profiles associated with air pollution. This may in part be explained by relatively small study sizes and modest exposure contrasts.

Considering the similarities between tobacco smoke (a combustion product) and air pollution (mostly combustion products), studies of smoking and gene expression might also provide some insight into potential gene expression targets of air pollution.[15] An example of such a study is Beineke et al.[16] in which 4,214 genes from biologic pathways known to be affected by both smoking and air pollution (i.e., apoptosis and cellular death, immune system development, leukocyte activation, hematopoiesis, stress response, and alterations in platelet activity) were correlated with self-reported smoking status.

In this hypothesis generating study, we assessed the association between annual average estimate of NO_x concentrations and genome-wide changes in gene expression in peripheral blood mononuclear cells in a large population using state-of-the-art exposure assessment methods.[17] We assessed the overlap between genes associated to NO_x and gene sets hypothesized to be related to air pollution exposure and cigarette smoking. Moreover, for genes for which expression levels were associated with NO_x, we assessed the role of DNA methylation in the regulation of gene expression at potentially cis-acting CpG sites. In addition, we assessed the interaction between the top-ranked probes and four inflammatory markers (*IL-2*, *IL-8*, *IL-10*, *TNF*-α) that we previously observed to be associated with NO_x in the same study population.[18]

METHODS

We combined data from two existing projects: gene expression profiles from the "Genomics Biomarkers of Environmental Health" (EnviroGenoMarkers)[19,20] and long-term average NO_x concentrations at residential addresses from the European Study of Cohorts for Air Pollution Effects (ESCAPE).[17] Study design and data collection procedures have been previously described in detail.[18]

Study Population

The EnviroGenoMarkers study was based on analyses of peripheral blood mononuclear cells of participants from two prospective cohorts: the Italian contribution to the European Prospective Investigation into Cancer and Nutrition study (EPIC-Italy, 95 individuals [22 men, and 73 women]) and the Northern Sweden Health and Disease Study (NSHDS, 455 individuals [202 men, and 253 women]). In both cohorts, blood samples were prospectively collected from healthy subjects at enrolment (around 1990–2006) and cohort members were asked to complete a standardized questionnaire focusing on dietary and life style.

Our study population, a subset of the EnviroGenoMarkers data, was collected in two phases and comprised in total 221 Non-Hodgkin's lymphoma cases and 58 breast cancer cases, identified through local cancer registries (loss to follow-up <2%), and the same number of controls matched on sex, age, center, and date of blood collection were included.[18] Cases were diagnosed on average 6 years (range 2–16 years) after recruitment/blood collection.

Ethics Statement

This study was approved by the committees on research ethics in Umea and Florence at the relevant institutions. All participants provided written consent at recruitment (EPIC-Italy 1993–1998; NSHDS 1990–2006).

Exposure Assessment

Annual modeled outdoor concentrations of NO_x at the study participant's home-address were available from the ESCAPE project.[17,21] Exposure to particulate matter (PM2.5, PM2.5 absorbance, and PM10) was only available for a subset of our study population (13 subjects). We therefore restricted our analyses to NO_x.[18] We natural-log-transformed the distribution of the NO_x concentration to limit the influence of high concentrations and to normalize the distribution.

Gene Expression Assessment

Total RNA was extracted from peripheral blood mononuclear cell samples stored within 2 hours of collection at −80°C. RNA from each sample was used to generate cDNA for array hybridization. The cDNA was then labeled with cyanine 3. The labeled cDNA was hybridized to Agilent whole human genome (4 × 44 K) arrays, containing 43,376 probes representing 29,846 genes. Subsequently, the hybridized slides were washed and scanned by using an Agilent Technologies G2565CA DNA Microarray scanner. Measurements for both phases were performed at Maastricht University. Technical performance and quality of the microarrays has been described in detail previously.[19,20] In short, microarray scan images were visually evaluated before and after within- and between-array normalization (LOESS and A-quantile, respectively). Good probes were identified based on the number of pixels, mean/median intensity ratio, saturation, or foreground/background intensity ratio. A total of 29,662 probes, representing 15,216 genes, were selected based on these criteria. We imputed missing values in Gene Pattern (version 3.1) using the *k* nearest neighbors approach ($k = 15$, Euclidian metric). When known, annotation of probes are provided in italics within parentheses.

Data Analysis

We performed univariate analyses to identify transcript concentration levels associated with long-term average exposure to NO_x. We complemented univariate analysis with two additional variable selection approaches (Elastic-Net regression and the Graphical Unit Evolutionary Stochastic Search Algorithm [GUESS]) that are capable of capturing the correlation among genes.[22] We call all probes that were identified in any of the statistical approaches "noteworthy probes."

Within the ESCAPE project,[17,21] the Swedish cohort was among the cohorts with the lowest levels of air pollution, while the Italian cohort was among the highest. We therefore stratified all statistical analyses by cohort (Table 1; Figure). As there was some overlap in the exposure distributions of the two cohorts, we also conducted analyses on the combined cohorts, while adjusting for country.

All statistical analyses were performed using R version 3.0.2 (packages: lme4,[23] glmnet,[24] c060,[25] and R2Guess[26]).

Univariate Mixed-effects Model

We conducted linear mixed-effects modeling to investigate the association between probe-level expression and long-term exposure to NO_x. To account for potential technical noise (nuisance variation), we incorporated the dates of three main steps of sample processing (i.e., RNA isolation, hybridization, and dye labeling) as random effects in the models. Exposure to NO_x and a priori selected potential confounding factors were included as fixed effects in the models. These confounding factors were body mass index (BMI) (kg/m^2), age (years) in three categories: (30–40, 41–50, 51–60), sex, smoking status (never smoker, former smoker, current smoker), phase (1 or 2), future disease status (lymphoma case, breast cancer case, control), and sample storage time (years), consistent with previous analyses of the EnviroGenoMarkers data.[19,27]

TABLE 1. Characteristics of the Study Population

Characteristics	Swedish Cohort (n = 455)	Italian Cohort (n = 95)
Sex (N [%])		
Female	253 (56)	73 (77)
Male	202 (44)	22 (23)
Smoking status (N [%])		
Current smoker	97 (21)	8 (8)
Former smoker	92 (20)	24 (25)
Never smoker	266 (59)	63 (66)
Age (years) (N [%])		
<40	28 (6)	2 (2)
40–50	134 (29)	29 (31)
>50	293 (65)	64(67)
Future disease status (N [%])[a]		
Breast cancer	46 (10)	12 (13)
Lymphoma	183 (40)	38 (40)
Control	226 (50)	45 (47)
BMI (kg/m^2) (mean ± SD)	26.1 ± 3.8	25.7 ± 3.7
NO_x (μg/m^3) (median [SD])	6.7 (5.8)	94.2 (42.5)

[a]Non-Hodgkin's lymphoma and breast cancer cases, identified through local Cancer Registries (loss to follow-up <2%), occurred on average 6 years (range 1–17 years) after recruitment/blood collection.

To assess how sensitive our findings were to variations in the confounder model, we conducted a set of additional analyses. We ran a minimally adjusted model (only age and sex included as covariates), a model in which smoking and BMI were excluded from the primary set of covariates, and a model in which we added education level (primary, technical, secondary, and university), and physical activity (moderately inactive, moderately active, and active) as covariates. In a further sensitivity analysis, we assessed the impact on our findings of adjusting our regression models for estimated cell-type composition[28] (available from an epigenome-wide analysis on a subset of the subjects).

We assessed the output from the univariate analysis using two approaches. First, we followed an agnostic approach using the Benjamini-Hochberg false discovery rate correction[29] to control for false positives. A false discovery rate <0.2 was used to classify probes as noteworthy.

Second, we followed a candidate gene approach, including 35 candidate genes from 10 biologic pathways relevant to air pollution exposure responses (coagulation, *Klf2*-mediated immune response, *NF-κB* signaling, acute phase response, *Nrf2*-mediated oxidative stress response, endoplasmic reticulum stress [*UPR*], glutathione metabolism, phase I and phase II metabolism, endogenous reactive oxygen species [*ROS*] production, and cytokine signaling)[13]—biologic candidate genes—augmented with genes that were associated to air pollution in epidemiologic studies (OMICS and gene–environment interaction) published since 2006 (empirical candidate genes; eTable 1; http://links.lww.com/EDE/B167). Empirical candidate genes were selected if an association in the same direction was observed in at least two previous studies. We assessed the overlap (strength and direction of the association) between the genes identified from the literature and results from our univariate analysis. In these analyses, we used a P value of 0.05 to classify probes as noteworthy for further evaluation.

Variable Selection Methods

Elastic-Net is a form of penalized multiple regression in which parameter estimates are achieved by using a combination of Ridge and Lasso penalties.[30] To control the number of falsely selected predictors by Elastic-Net, we applied the stability selection technique proposed by Meinshausen and Bühlmann.[31] We accounted for multiple testing by setting family-wise error rate (FWER) to 0.05. As a sensitivity analysis, we also set the family-wise error rate at 0.2, relaxing the type one error. We set the threshold of selection probability (probability of selecting a predictor by algorithm; π) to 0.6.

GUESS is a Bayesian variable selection approach that uses an advanced stochastic search Markov Chain Monte Carlo algorithm.[32] GUESS fits a range of models containing

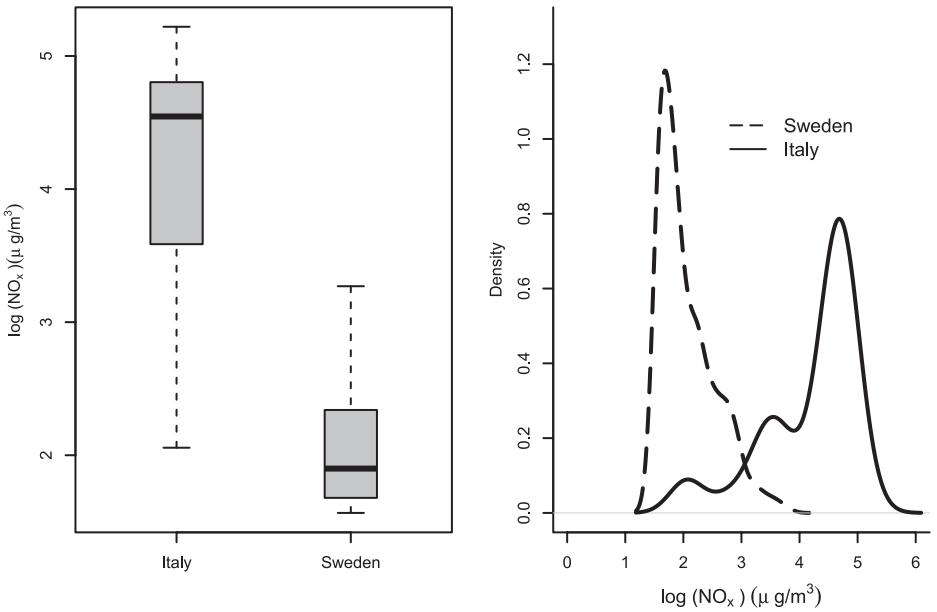

FIGURE. Box plot (left) and density plot (right) of log(NO_x) (μg/m^3) concentration for Swedish and Italian participants. Log(NO_x) concentrations are shown on the *Y* axis of the box plots and on the *X* axis of the density plot.

varying combinations of probes (predictors) to the data. Noteworthy probes were selected based on the marginal posterior probability of inclusion, which provides a model-averaged measure of importance of each probe with respect to the models that were fit to the data. We ran five different chains in GUESS for 90k iterations and discarded the first 30k iterations as burn-in. Expected and SD of the model size were set to 3 and 5, respectively. We used a marginal posterior probability of inclusion of 10% to call a probe noteworthy (technical details in eAppendix 1; http://links.lww.com/EDE/B167).

DNA Methylation at Relevant Loci

We assessed the association between long-term exposure to NO_x and degree of methylation for all CpG islands (n = 74) that were in cis position to the noteworthy genes identified in the gene expression analysis. Methylation data generated using the Infinium HumanMethylation450 BeadChip (450K) was available for a subset (466 of 550) subjects.[28] We used the same univariate mixed-effects models as described above for our main gene expression analyses.

Correlation Between Noteworthy Probes and Inflammatory Plasma Markers

To explore whether our noteworthy transcripts and markers of inflammation in plasma (IL-2, IL-8, IL-10, TNF-α), which we previously observed to be associated to air pollution in this cohort,[18] potentially operate in similar biologic pathways, we assessed the interrelationships (Pearson correlation) between identified transcripts and inflammatory markers.

Gene Set Enrichment Analysis

We used gene set enrichment analysis[33] to assess whether the gene expression pattern associated with NO_x in our data has similarities to patterns associated with cigarette smoking or air pollution responses. We assessed whether the distribution of overlap between two sets of genes—associated with either "exposure to cigarette smoke" or "biologic responses due to exposure to air pollution"—and all genes included in our study was random, or whether this overlap primarily occurred among our most up- or downregulated genes (i.e., enrichment). The first gene set comprised 4,214 genes whose expression in peripheral blood was associated with smoking status in a study by Beineke et al.[16] (smoking set). The second gene set comprised 35 genes from 10 biological pathways relevant to air pollution responses[13] (air pollution set). A similar approach using gene set enrichment analysis was described by Wang et al.[34] demonstrating enrichment of cigarette smoke-related[34] genes among genes affected by indoor air pollution (technical details in eAppendix 2; http://links.lww.com/EDE/B167).

To enhance interpretability, we conducted gene set enrichment analysis separately for upregulated genes (all genes with positive *t* statistic in our univariate analysis) and downregulated genes (all genes with negative *t* statistic). We compared our upregulated genes with the upregulated genes in the smoking and air pollution sets and compared our downregulated genes with the downregulated genes in the smoking set (no downregulated genes were included in the air pollution set). A *P* value <0.05 was used as statistical cut-off for enrichment.

RESULTS

Table 1 summarizes the baseline characteristics of the study participants. The Swedish cohort has a lower proportion of women than the Italian cohort (56% vs. 77%), a higher proportion of current smokers (21% vs. 8%), and a lower proportion of never smokers (59% vs. 66%).

We observed a considerable difference in the distribution of NO_x concentrations between the two countries

(Table 1; Figure). The median (5th percentile, 95th percentile) concentration of NO_x estimated for the Italian cohort (94.2 μg/m^3 [7.8, 124.6]) was considerably higher than the median concentration estimated for Sweden (6.7 μg/m^3 [4.8, 19.5]).

Univariate Mixed-effects Model

Agnostic Approach

Following an agnostic approach, we identified six noteworthy probes that were associated (false discovery rate <0.2) to long-term average exposure to NO_x in the Italian cohort (Table 2). These probes are A_23_P252075 (*AHCYL2*) (*q* value 0.12), A_24_P406830 (*MTMR2*) (*q* value 0.12), A_32_P175313 (*q* value 0.17), A_32_P44961 (*LARP1B*) (*q* value 0.02), A_32_P156373 (*q* value 0.17), and A_32_P61298 (*q* value 0.17). In Table 2, we show the results from the Swedish and combined cohorts for the top-ranked probes (based on the *q* value) in the Italian cohort. Perturbations of the noteworthy probes in Italy were all in the same direction in the combined cohort and for three probes (A_23_P252075 (*AHCYL2*), A_32_P44961 (*LARP1B*), A_32_P61298) in the Swedish cohort. However, these differences did not reach the threshold (BH-FDR < 0.2) in either the Swedish or the combined cohort.

To formally explore heterogeneity between cohorts, we conducted analyses in the combined population, while including an interaction term between cohort and NO_x. We observed an interaction for four out of six noteworthy probes that were associated to NO_x in the Italian cohort, but not in the Swedish cohort (Table 2) (complete results [by cohort and combined] are available in eAppendix3.DOI10.5281/zenodo.50661; http://links.lww.com/EDE/B167).

Sensitivity Analyses

To assess how sensitive our findings were to variations in the confounder model, we conducted a set of additional analyses. Applying a model that was only adjusted for sex and age, resulted in eight additional probes being associated (BH-FDR < 0.2) with NO_x in the Italian cohort. One of the original findings, probe A_32_P61298, was not retained in these analyses. Applying a model in which smoking and BMI were excluded as covariates, four of the six probes identified in our primary univariate analyses were retained (all except A_32_P175313 and A_32_P61298) and four additional probes were associated (BH-FDR < 0.2) with NO_x in the Italian cohort. Applying a model in which we added education and physical activity as covariates, we observed three probes (A_32_P44961 (*LARP1B*), A_32_P175313, A_23_P252075 (*AHCYL2*)) to be associated (BH-FDR < 0.2) with NO_x in the Italian cohort. All three probes were included in our primary model. Applying a model in which we corrected for cell-type composition resulted in 10 additional probes being associated (BH-FDR < 0.2) to NO_x in the Italian cohort. One of the original findings, probe A_32_P175313 (*q* value = 0.226), no longer met the cut-off (BH-FDR < 0.2) for noteworthiness after correction for cell-type composition.

Candidate Gene Approach

We included 36 probes in our candidate gene approach. Eight probes were selected on empirical grounds and 30 probes (corresponding to 26 genes that are overlapping between our study and study by Wittkopp et al.[13]) were selected based on biologic motivation. *IL-6* and *HMOX1* were included in both lists of candidate genes. Studies from which these genes were selected are listed in eTable 1 (http://links.lww.com/EDE/B167). We present parameter estimates and associated *P* values from univariate mixed-effects regression in Table 3. We observed a positive association of NO_x with *NOX1* in the Italian cohort and with *IL-8* in the Swedish (and combined) cohort. The direction of these effects was in agreement with what has been reported in the literature. One gene was negatively

TABLE 2. Selected Associations Between Long-term Exposure to NO_x and Transcript Levels Based on Agnostic Approach (*q* Value <0.2) in the Italian, Swedish, and Combined Population

		Italian Cohort		Swedish Cohort		Combined Cohort		
Agilent ID	Gene Name	β[a] (SE)	*Q* Value[b]	β (SE)	*Q* Value	β (SE)	*Q* Value	*P* Value for Interaction[c]
A_23_P252075	*AHCYL2*[d]	0.23 (0.05)	0.12	0.01 (0.04)	0.99	0.10 (0.03)	0.99	3×10^{-4}
A_24_P406830	*MTMR2*[e]	−0.20 (0.05)	0.12	0.02 (0.03)	0.99	−0.05 (0.03)	0.99	2×10^{-4}
A_32_P156373	Unknown	0.31 (0.07)	0.17	−0.06 (0.07)	0.99	0.06 (0.06)	0.99	3×10^{-4}
A_32_P175313	Unknown	0.30 (0.07)	0.17	−0.05 (0.06)	0.99	0.07 (0.05)	0.99	1×10^{-3}
A_32_P44961	*LARP1B*[f]	0.33 (0.07)	0.02	0.04 (0.06)	0.99	0.13 (0.05)	0.99	6×10^{-1}
A_32_P61298	Unknown	0.38 (0.09)	0.17	0.01 (0.09)	0.99	0.07 (0.07)	0.99	8×10^{-1}

[a]Effect estimate per unit changes of the exposure.
[b]*Q* value, false discovery rate correction for *P* value.
[c]*P* value, for the interaction between country and NO_x in the combined population.
[d]Adenosylhomocysteinase-like2.
[e]Myotubularin-related protein 2.
[f]La ribonucleoprotein domain family member 1B.
SE indicates standard error.

TABLE 3. Parameter Estimates and Standard Error from Univariate Regression for Candidate Genes (Both Empiric and Biologic) Previously Associated with Air Pollution in the Epidemiologic Literature

		Italy Cohort	Swedish Cohort	Combined Cohort
Agilent ID	**Gene Name**	**β[a] (SE)**	**β (SE)**	**β (SE)**
Empirical candidate genes				
A_23_P502464	*NOS2A*[b]	−0.03 (0.08)	0.03 (0.06)	0.02 (0.05)
A_32_P50123	*SRGAP2*	−0.16 (0.08)	0.01 (0.08)	−0.06 (0.06)
A_24_P357869		−0.01 (0.07)	−0.08 (0.05)	−0.06 (0.04)
A_23_P200829		−0.002 (0.06)	0.04 (0.04)	0.02 (0.03)
A_23_P115407	*GSTM1*	−0.06 (0.08)	−0.05 (0.07)	−0.03 (0.04)
A_23_P202658	*GSTP1*	−0.084 (0.08)	0.04 (0.06)	−0.01 (0.05)
Empirical and biological candidate genes				
A_23_P120883	*HMOX1*	0.07 (0.08)	0.07 (0.06)	0.06 (0.05)
A_23_P71037	*IL-6*	0.004 (0.14)	0.11 (0.10)	0.05 (0.08)
Biological candidate genes				
A_23_P103996	*GCLM*	0.06 (0.06)	−0.04 (0.04)	−0.01 (0.03)
A_32_P177953		0.06 (0.06)	−0.04 (0.05)	−0.003 (0.04)
A_23_P105138	*CAT*	0.01 (0.06)	0.01 (0.05)	0.01 (0.04)
A_23_P119196	*KLF2*	0.04 (0.08)	0.04 (0.07)	0.05 (0.05)
A_24_P151305		−0.02 (0.05)	−0.04 (0.05)	−0.01 (0.04)
A_23_P120933	*ATF4*	−0.10 (0.07)	−0.05 (0.05)	−0.07 (0.04)
A_23_P120941		−0.01 (0.04)	−0.07 (0.04)	−0.05 (0.03)
A_23_P137697	*SELP*	−0.15 (0.11)	−0.12 (0.08)	−0.13 (0.07)
A_23_P154840	*SOD1*	−0.01 (0.06)	0.02 (0.04)	0.02 (0.03)
A_23_P163402	*CYP1A1*	0.10 (0.08)	0.08 (0.07)	0.09 (0.05)
A_23_P202658	*GSTP1*	−0.08 (0.08)	0.04 (0.06)	−0.01 (0.05)
A_23_P204581	*TXNRD1*	−0.01 (0.06)	−0.09 (0.11)	−0.08 (0.08)
A_23_P209625	*CYP1B1*	0.13 (0.11)	0.08 (0.07)	0.07 (0.06)
A_23_P215566	*AHR*	−0.01 (0.11)	0.03 (0.06)	−0.002 (0.05)
A_23_P217280	*NOX1*	0.26 (0.13)	0.05 (0.11)	0.10 (0.09)
A_23_P250671	*GPX1*	−0.01 (0.05)	0.02 (0.04)	0.001 (0.03)
A_23_P352879	*GCLC*	0.11 (0.07)	0.01 (0.05)	0.04 (0.04)
A_23_P5761	*NFE2L2*	−0.01 (0.06)	−0.05 (0.04)	−0.04 (0.04)
A_23_P62907	*ATF6*	0.09 (0.07)	0.01 (0.05)	0.03 (0.04)
A_23_P7144	*CXCL1*	−0.18 (0.13)	−0.13 (0.13)	−0.16 (0.10)
A_23_P79518	*IL1B*	−0.06 (0.18)	−0.13 (0.14)	−0.05 (0.12)
A_23_P89380	*IL-8*	0.11 (0.11)	0.25 (0.09)	0.23 (0.07)
A_23_P89431	*CCL2*	0.13 (0.11)	0.01 (0.09)	0.05 (0.07)
A_24_P379413	*IL6R*	−0.02 (0.08)	−0.03 (0.06)	−0.02 (0.05)
A_24_P77008	*PTGS2*	−0.24 (0.18)	−0.13 (0.16)	−0.17 (0.12)
A_24_P935819	*SOD2*	−0.21 (0.12)	−0.11 (0.13)	−0.13 (0.10)
A_24_P936444	*NFE2L2*	0.05 (0.06)	−0.08 (0.04)	−0.02 (0.04)
A_32_P13728	*HSPA8*	−0.02 (0.09)	0.046 (0.05)	0.03 (0.04)

[a]Effect estimate per unit changes of the exposure. Gene abbreviations are listed in eTable (http://links.lww.com/EDE/B167).
SE indicates standard error.

associated with exposure to NO_x in our analysis (*SELP* in the combined cohort), but the direction of this effect was not in agreement with what has been reported in the literature.

Elastic-Net

Application of Elastic-Net with stability selection yielded one probe (A_32_P44961 [*LARP1B*]) that was associated (FWER < 0.05) with long-term average exposure to NO_x in the Italian cohort. By increasing the FWER cut-off to 20%, Elastic-Net selected three more probes (A_32_P156373, A_32_P175313, and A_23_P252075 [*AHCYL2*]). All probes were also identified using univariate analysis. We observed no evidence for an association between long-term exposure to NO_x and gene expression in the combined and Swedish cohort for either FWER cut-off values.

GUESS

GUESS did not identify any probe that exceeded our predefined MPPI cut-off level of 10%.

DNA Methylation at Relevant Loci

We assessed the association between long-term exposure to NO_x and degree of methylation for 74 CpG islands in cis with our five noteworthy genes. Methylation data were available for a subset (466 of 550) subjects. Results from this analysis are presented in eTable 2 (http://links.lww.com/EDE/B167). Methylation of two CpG islands (hypomethylation cg03793937 upstream of *MTMR2*; q value = 0.14, and hypermethylation cg06988775 downstream of *AHCYL2*; q value = 0.14) was associated (false discovery rate < 0.2) with long-term exposure to NO_x in the Italian cohort.

Correlation Between Noteworthy Probes and Inflammatory Markers

The Pearson correlation coefficients between concentrations of four inflammatory markers from our previous study[18] (i.e., *IL-2*, *IL-8*, *IL-10*, and *TNF*-α) and expression levels of the eight noteworthy probes from the agnostic and candidate gene approaches are presented in Table 4 for Swedish and Italian cohorts, separately.

The overall median of correlation coefficients and their interquartile range for Italian and Swedish cohorts are = −0.04 (−0.05, 0.11) and = −0.01 (−0.02, 0.03), respectively. The highest correlation was observed between *IL-2* and two of the noteworthy probes A_24_P406830 (*MTMR2*) (= 0.4) and A_32_P175313 (= 0.28) in the Italian cohort.

Gene Set Enrichment Analysis

We conducted gene set enrichment analysis using the results for the Italian cohort. When we compared our results with the set of genes negatively associated to cigarette smoking status, we observed enrichment of the overlapping genes among the genes that were most strongly downregulated due to exposure to air pollution in our analysis (eFigure 1A; http://links.lww.com/EDE/B167). We did not observe enrichment when we compared our upregulated genes to the upregulated smoking genes (eFigure 1B; http://links.lww.com/EDE/B167) or the air pollution gene set (eFigure 2; http://links.lww.com/EDE/B167).

DISCUSSION

Our study provides some evidence that subtle changes in gene expression are associated with long-term exposure to air pollution as measured by NO_x in a cohort of adult individuals. We identified seven noteworthy probes in the analysis of the Italian cohort, and one noteworthy probe in the analysis of the Swedish (and the combined Italian-Swedish) cohort. Of these A_23_P252075 (*AHCYL2*) and A_32_P44961 (*LARP1B*) achieved BH_FDR < 0.2 in all sensitivity analyses. Gene-set enrichment analyses indicated that our downregulated genes overlapped to a certain extent with genes for which smoking-induced gene expression differences were previously published.

All noteworthy probes we identified in our agnostic analysis are novel findings. Targeted analyses provided support for a potential association with *NOX1 and IL-8* and air pollution, whereas for some genes (e.g., *Nrf2*-mediated genes *HMOX1*, *NQO1*, and *SOD2*[13]), our point estimates suggested associations in the same direction as previously reported, but these associations were imprecise.

To date, few reports of associations between air pollution and gene expression have been replicated. We attribute the lack of replication to the small study populations that have been used, the relatively low exposure levels that individuals in these cohorts experienced, and the lack of adjustment for multiple testing in most studies.[6,7,13,14] Although our study population is still modest in size, the 95 individuals in the Italian cohort were among the highest exposed within Europe.[21]

In studies such as ours, the risk of false-positive findings is high due to the large number of tests conducted compared with the number of observations available. For the univariate analysis, we reduced this risk by controlling the false discovery rate and by using evidence from the literature as an informal prior in our analysis. A limitation of univariate

TABLE 4. Correlation Between Expression Levels of Noteworthy Probes and Concentrations of Inflammatory Markers Previously Reported to be Associated with NO_x in the Same Study Population

		Italy				Sweden			
Noteworthy Probes	**Gene Name**	**IL.2**	**IL.8**	**IL.10**	**TNF-α**	**IL.2**	**IL.8**	**IL.10**	**TNF-α**
A_23_P252075	*AHCYL2*	−0.18	−0.05	−0.04	−0.04	0.007	0.018	0.057	−0.004
A_24_P406830	*MTMR2*	0.40	0.03	0.00	0.18	−0.082	0.030	−0.044	−0.067
A_32_P156373	Unknown	−0.10	0.09	0.09	−0.01	0.055	−0.015	0.063	−0.007
A_32_P175313	Unknown	0.28	0.16	0.15	0.20	0.001	0.071	0.041	−0.021
A_32_P44961	*LARP1B*	−0.05	0.07	0.06	−0.08	0.002	−0.004	−0.010	−0.067
A_32_P61298	Unknown	0.00	−0.06	−0.13	−0.13	−0.017	0.000	−0.007	−0.068
A_23_P217280	*NOX1*	0.053	−0.12	−0.18	−0.11	−0.018	0.058	0.014	−0.019
A_23_P89380	*IL-8*	0.070	0.126	0.113	0.073	0.015	0.093	0.013	0.049

analysis is that it cannot take the correlation structure among genes into account, further increasing the risk of false-positive findings.[22] We therefore applied two additional approaches (Elastic-Net and GUESS) that are capable of capturing the correlation among genes. Application of these approaches reduces the risk of false-positive findings, but at the cost of reduced sensitivity.[35] Correspondingly, in the present study, agnostic univariate analysis identified six noteworthy genes that were associated with NO_x, Elastic-Net regression (FWER 0.2) selected four of these whereas GUESS did not select any. We view the three approaches as complementary, but as the primary goal of our current analysis was to discover potential new gene expression targets of ambient air pollution, we preferred high sensitivity over a lower risk of false-positive findings.

Heterogeneity in Results Across Cohorts

As hypothesized, we observed stronger signals in the Italian cohort than in the Swedish cohort. This is likely due to the different level of exposure in the two cohorts. Although the exposure assessment strategy in both cohorts was the same, absolute exposure levels and the exposure contrast in the Swedish cohort were low (median 6.65 μg/m^3, SD 5.8), compared with the Italian cohort (median 94.21 μg/m^3, SD 43.0).

Biological Role of Noteworthy Genes

Using the Gene Expression Omnibus,[36] we identified gene annotations for five of our noteworthy probes (*AHCYL2*, *MTMR2*, *LARP1B, IL-8*, and *NOX1*). Functional analysis of these genes using the NIH-DAVID bioinformatics resources[37] yielded no evidence for functional enrichment in any biologic pathway likely due to the limited number of probes found in this study. Biological roles of noteworthy genes based on literature review are presented in eTable 3 (http://links.lww.com/EDE/B167).

Gene Set Enrichment Analysis

Using our univariate and variable selection approaches, we focused on a small set of genes that showed the largest perturbation. However, following this approach, we might have missed signals that did not meet our threshold for noteworthiness because the perturbation in gene expression was modest relative to the noise inherent to the microarray technology.[33] Gene set enrichment analysis overcomes this limitation by using information (ranking according to the strength of the association with exposure to air pollution) from multiple genes rather than assessing the genes one by one.[33] The observation that there was enrichment of smoking-associated genes among the genes that were negatively associated to exposure to NO_x in our study provides some indication that the perturbation of the transcriptome by exposure to air pollution we observed was a true finding, rather than a false positive. Furthermore, this observation points toward a shared biologic pathway of the effects of cigarette smoke and air pollution on the transcriptome which is of interest due to overlap between health outcomes that have been related to tobacco smoking and air pollution.[34]

DNA Methylation at Relevant Loci

For two genes (*MTMR2* and *AHCYL2*), we observed an effect of long-term exposure to NO_x on expression level and on methylation status of two CpG islands in cis with these genes. The effects of long-term exposure to NO_x on methylation of the two genes were in the same direction as what we observed for gene expression (downregulation for *MTMR2* and upregulation for *AHCYL2*), which does not confirm to the often observed inverse correlation between methylation in promoter regions and gene expression due to a gene silencing effect of methylation. In addition, as we tested 74 CPG islands, there is a likelihood for false positives and we are therefore cautious in interpreting these results as a cross-OMICS signal of air pollution.

Correlation Between Noteworthy Probes and Inflammatory Markers

We found positive and relatively high correlation between immune marker *IL-2* and two noteworthy probes (A_24_P406830 [*MTMR2*], A_32_P175313) in the Italian cohort but not in the Swedish cohort. We did not identify further information regarding a potential shared pathway between the genes and the immune marker. The fact that we observed a correlation in the Italian cohort, but not in the Swedish cohort does provide some indication for a potential role of NO_x in inducing this correlation in the Italian cohort by affecting both probes; A_24_P406830 (*MTMR2*), A_32_P175313 and the concentration of *IL-2*.

A limitation of our study is that we have included probes without knowing where they bind to the genome (nonannotated probes). Although results for nonannotated probes are less informative than the probes that have been annotated, they do provide some indication of general perturbation of the transcriptome by air pollution. To assess the impact of this decision, we removed the 6,848 unannotated probes from the analyses. These analyses did not identify any newly associated probes.

CONCLUSION

In summary, our study provides some evidence for subtle changes in the transcriptome in relation to long-term exposure to NO_x. Some of these changes are consistent with transcriptome perturbations that have been observed among tobacco smokers. Our results contribute to the further elucidation of the pathways through which long-term exposure to air pollution induces adverse health effects.

ACKNOWLEDGMENTS

We greatly acknowledge all those who are responsible for data management in both the European Study of Cohorts for Air Pollution Effects and EnviroGenoMarkers cohorts.

REFERENCES

1. Brunekreef B, Holgate ST. Air pollution and health. *Lancet*. 2002;360:1233–1242.
2. Hoek G, Krishnan RM, Beelen R, et al. Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environ Health*. 2013;12:43.
3. Ghio AJ, Carraway MS, Madden MC. Composition of air pollution particles and oxidative stress in cells, tissues, and living systems. *J Toxicol Environ Health B Crit Rev*. 2012;15:1–21.
4. Li N, Hao M, Phalen RF, Hinds WC, Nel AE. Particulate air pollutants and asthma. A paradigm for the role of oxidative stress in PM-induced adverse health effects. *Clin Immunol*. 2003;109:250–265.
5. Demetriou CA, Raaschou-Nielsen O, Loft S, et al. Biomarkers of ambient air pollution and lung cancer: a systematic review. *Occup Environ Med*. 2012;69:619–627.
6. Huang YC, Schmitt M, Yang Z, et al. Gene expression profile in circulating mononuclear cells after exposure to ultrafine carbon particles. *Inhal Toxicol*. 2010;22:835–846.
7. Pettit AP, Brooks A, Laumbach R, et al. Alteration of peripheral blood monocyte gene expression in humans following diesel exhaust inhalation. *Inhal Toxicol*. 2012;24:172–181.
8. Li N, Alam J, Venkatesan MI, et al. Nrf2 is a key transcription factor that regulates antioxidant defense in macrophages and epithelial cells: protecting against the proinflammatory and oxidizing effects of diesel exhaust chemicals. *J Immunol.* 2004;173:3467–3481.
9. Huang Y-CT, Karoly ED, Dailey LA, et al. Comparison of gene expression profiles induced by coarse, fine, and ultrafine particulate matter. *J Toxicol Environ Health A*. 2011;74:296–312.
10. Araujo JA, Barajas B, Kleinman M, et al. Ambient particulate pollutants in the ultrafine range promote early atherosclerosis and systemic oxidative stress. *Circ Res*. 2008;102:589–596.
11. van Leeuwen DM, Gottschalk RW, Schoeters G, et al. Transcriptome analysis in peripheral blood of humans exposed to environmental carcinogens: a promising new biomarker in environmental health studies. *Environ Health Perspect*. 2008;116:1519–1525.
12. van Leeuwen DM, van Herwijnen MH, Pedersen M, et al. Genome-wide differential gene expression in children exposed to air pollution in the Czech Republic. *Mutat Res*. 2006;600:12–22.
13. Wittkopp S, Staimer N, Tjoa T, et al. Nrf2-related gene expression and exposure to traffic-related air pollution in elderly subjects with cardiovascular disease: an exploratory panel study. *J Expo Sci Environ Epidemiol*. 2016;26:141–149.
14. Peretz A, Peck EC, Bammler TK, et al. Diesel exhaust inhalation and assessment of peripheral blood mononuclear cell gene transcription effects: an exploratory study of healthy human volunteers. *Inhal Toxicol*. 2007;19:1107–1119.
15. Pope CA 3rd, Burnett RT, Turner MC, et al. Lung cancer and cardiovascular disease mortality associated with ambient air pollution and cigarette smoke: shape of the exposure-response relationships. *Environ Health Perspect*. 2011;119:1616–1621.
16. Beineke P, Fitch K, Tao H, et al.; PREDICT Investigators. A whole blood gene expression-based signature for smoking status. *BMC Med Genomics*. 2012;5:58.
17. Beelen R, Hoek G, Vienneau D, et al. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – The ESCAPE project. *Atmos Environ*. 2013;72:10–23.
18. Mostafavi N, Vlaanderen J, Chadeau-Hyam M, et al. Inflammatory markers in relation to long-term air pollution. *Environ Int*. 2015;81:1–7.
19. Chadeau-Hyam M, Vermeulen RC, Hebels DG, et al.; EnviroGenoMarkers project consortium. Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Ann Oncol*. 2014;25:1065–1072.
20. Hebels DG, Georgiadis P, Keun HC, et al.; EnviroGenomarkers Project Consortium. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. *Environ Health Perspect*. 2013;121:480–487.
21. Cyrys J, Eeftens M, Heinrich J, et al. Variation of NO2 and NOx concentrations between and within 36 European study areas: results from the ESCAPE study. *Atmos Environ*. 2012;62:374–390.
22. Chadeau-Hyam M, Campanella G, Jombart T, et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen*. 2013;54:542–557.
23. Boeck P De, Bakker M, Zwitser R, et al. The estimation of item response models with the lmer function from the lme4 Package in R. *J Stat Sof tw.*2011;39:1–28.
24. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.*2010;33:1–22.
25. Sill M, Hielscher T, Becker N, Zucknick M. c060 : Extended inference with lasso and elastic-net regularized cox and generalized linear models.*J Stat Softw*. 2014;62:1–22.
26. Liquet B, Bottolo L, Campanella G, Richardson S, Chadeau-Hyam M. R2GUESS : a graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses. *J Stat Softw*. 2016;69:1–32.
27. Kelly RS, Lundh T, Porta M, et al.; EnviroGenoMarkersProject Consortium. Blood erythrocyte concentrations of cadmium and lead and the risk of B-cell non-Hodgkin's lymphoma and multiple myeloma: a nested case-control study. *PLoS One*. 2013;8:e81892.
28. Georgiadis P, Hebels DG, Valavanis I, et al.; EnviroGenomarkers consortium. Omics for prediction of environmental health effects: blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. *Sci Rep*. 2016;6:20544.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 1995;57:289–300.
30. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B.* 2005;67:301–320.
31. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B.* 2010;72:417–473.
32. Bottolo L, Chadeau-Hyam M, Hastie DI, et al. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet*. 2013;9:e1003657.
33. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–15550.
34. Wang TW, Vermeulen RC, Hu W, et al. Gene-expression profiling of buccal epithelium among non-smoking women exposed to household air pollution from smoky coal. *Carcinogenesis*. 2015;36:1494–1501.
35. Agier L, Portengen L, Chadeau-Hyam M, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ Health Perspect*. 2016;124:1848–1856.
36. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*. 2013;41(Database issue):D991–D995.
37. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44–57.